A reason-based theory of rational choice

Franz Dietrich, Christian List

A reason-based theory of rational choice

RM/09/057

METEOR

Maastricht University School of Business and Economics
Maastricht Research School of Economics
of Technology and Organization

P.O. Box 616
NL - 6200 MD Maastricht
The Netherlands

A reason-based theory of rational choice

Franz Dietrich and Christian List
London School of Economics

October 29, 2009

1 Introduction

The idea that a rational choice is a choice based on reasons and that a rational agent is someone who acts on the basis of reasons – perhaps the right reasons – is a very natural one, and yet reasons are largely absent from modern rational choice theory. Instead, rational choice theory, also known as decision theory, is thoroughly Humean.[1] A rational agent, on the standard picture, has beliefs (typically represented by a subjective probability function over possible worlds or states of the world) and desires (typically represented by a utility function over the same or the outcomes of actions in them),[2] and acts so as to satisfy his or her desires maximally in accordance with his or her beliefs. In particular, the agent's desires over possible worlds or fully specified outcomes, also called *fundamental preferences*, are emotive attitudes and are entirely separate from and unresponsive to his or her beliefs, which are cognitive attitudes. Only surface-level desires, or *derived preferences*, over non-fundamental prospects such as sets of possible worlds can change in response to changes in beliefs about the relative likelihood

[1] For classic contributions, see, among many others, John von Neumann and Oskar Morgenstern, *Theory of Games and Economic Behavior* (Princeton: Princeton University Press, 1944), Leonard Savage, *The Foundations of Statistics* (New York: Wiley, 1954), and Richard Jeffrey, *The Logic of Decision* (Chicago: University of Chicago Press, 1965/1983).

[2] To be precise, in Savage's theory probabilities are defined over possible states of the world, and utilities over possible outcomes of actions in them, whereas in Jeffrey's theory both probabilities and utilities are defined over possible worlds.

of different outcomes of those prospects. Fundamental preferences, however, are fixed and outside the realm of rational – especially reason-based – scrutiny.

The aim of this paper is to present an alternative theory of rational choice, which gives reasons its proper place. Of course, there is a substantial body of philosophical work on the relationship between reasons and actions, but there is currently no *formal* theory of rational choice that is reason-based.[3] As a result, decision theorists and social scientists engaged in the formal modelling of decision problems do not have at their disposal the conceptual resources for capturing the role played by reasons in rational decision making.[4] This is an important gap in the literature, which we here seek to fill.

We develop a theory of rational choice in which an agent's preferences – and subsequently his or her choices and actions – are explained by the set of reasons that motivate him or her. The relationship between motivating reasons and preferences is described by two axioms, whose consequence is a general representation of the agent's preferences in terms of an underlying binary relation over the different possible sets of reasons that may be instantiated by the objects of preference. This binary relation, to be called the agent's *generating relation*, can be interpreted in a number of ways. On a 'cognitivist' interpretation, it encodes a particular set of propositions about

[3] A review of this rich philosophical literature is beyond the scope of this paper. Important contributions include Derek Parfit, *Reasons and Persons* (New York: Oxford University Press, 1984), Thomas M. Scanlon, *What We Owe to Each Other* (Cambridge, MA: Harvard University Press, 1998), Joseph Raz, *Practical Reason and Norms* (Oxford: Oxford University Press, 1999), and the essays in R. Jay Wallace, Philip Pettit, Samuel Scheffler and Michael Smith (eds.), *Reason and value: themes from the moral philosophy of Joseph Raz* (Oxford: Oxford University Press, 2004), notably John Broome's chapter, 'Reasons', pp. 28-55. See also the recent review articles by Stephen Finlay and Mark Schroeder, 'Reasons for Action: Internal vs. External', *Stanford Encyclopedia of Philosophy* (2008), and James Lenman, 'Reasons for Action: Justification vs. Explanation', *Stanford Encyclopedia of Philosophy* (2009), both available at: http://plato.stanford.edu/

[4] In *Reason and Rationality* (Princeton: Princeton University Press, 2009), Jon Elster confirms this observation: 'Whereas the theory of rational choice has been elaborated and developed with great precision, the same cannot be said of the idea of reason' (p. 7).

the relative 'goodness' of different sets of reasons, while on a 'non-cognitivist' interpretation, it encodes the agent's dispositions to prefer certain sets of reasons to others. On either interpretation, the agent's generating relation can be seen as the fundamental basis which, together with his or her motivating reasons, induces his or her preferences.

Our theory is both ambitious and flexible. It not only captures the idea that a rational choice is a choice based on reasons, and that a rational agent is someone whose actions are motivated by reasons, but it also shows us how changes in the set of reasons motivating an agent can lead to changes in the agent's preferences, even at the level of possible worlds or fully specified outcomes, where standard rational choice theory takes preferences to be fundamental and unchangeable. By implication, our theory has powerful resources for illuminating the relationship between rational choice and deliberation about reasons, which in turn is relevant to many theoretical and practical questions in philosophy and the social sciences. In addition, our theory generalizes standard rational choice theory, entailing it as a special case, and thereby pinpoints precisely what restriction the absence of reasons from a theory of rational choice implies.[5]

The paper is structured as follows. We begin by introducing our basic concepts – alternatives, preferences and reasons – and discuss what it means for a reason to be motivating (sections 2 and 3). Next, after explaining how our theory depicts an agent's possible psychological states, we introduce two axioms on the relationship between reasons and preferences (sections 4 and 5). We then present our two main representation theorems and discuss their interpretation (sections 6, 7 and 8). With this core of the theory in

[5] In their recent working paper, 'Rationalization' (University of Pennsylvania, 2008), Vadim Cherepanov, Timothy Feddersen and Alvaro Sandroni present a generalization of standard rational choice theory that can account for what they call the phenomenon of 'rationalization'. The central idea is that, although agents have fixed underlying preferences, they are constrained to make choices that they can 'rationalize', for instance justify in public. Although insightful in many ways, their model remains close to standard rational choice theory, in so far as the assumption of fixed fundamental preferences is not challenged. Some limitations of standard rational choice theory are discussed in Philip Pettit, 'Decision Theory and Folk Psychology', in Michael Bacharach and Susan Hurley (eds.), *Foundations of Decision Theory* (Oxford: Blackwell, 1991), pp. 147-175.

place, we comment on some philosophical questions, first on the distinction between explanatory and justificatory reasons and secondly on the role of reasons in an agent's rational deliberation and on their relevance to the resolution of disagreements between different agents' preferences (sections 9 and 10). Finally, we add two extensions without which our theory would not be complete: We show how it can handle preferences under uncertainty, and explain how an agent's reason-based preferences lead to his or her choices (sections 11 and 12). Formal proofs are given in an appendix.

2 Alternatives, preferences and reasons

We consider an agent's preferences over some set of fundamental objects of preference, which we call *alternatives*. Depending on the area of application, the alternatives could be, for example, possible worlds or states of the world, outcomes of actions, policy programmes, bundles of goods, or election candidates. What matters is that those alternatives are mutually exclusive and jointly exhaustive of the relevant space of possibilities. Later we also consider preferences over general prospects, that is, probability distributions over alternatives, so as to capture the fact that agents often cannot choose between individual alternatives, but only between different uncertain prospects, which are induced by the actions the agent can take.

Let X denote the set of all alternatives. A subset of X is called a proposition; it is said to be *true* of those alternatives contained in it, and *false* of all others. The agent's preferences over the alternatives in X are represented by some order $\succsim$ (a complete and transitive binary relation) on X. For any two alternatives x and y, we write $x \succsim y$ to mean that the agent weakly prefers x to y. We further write $x \succ y$ if $x \succsim y$ but not $y \succsim x$ (the case of a strict preference for x over y), and $x \sim y$ if $x \succsim y$ and $y \succsim x$ (the case of an indifference between x and y). Later we make explicit the way in which these preferences affect the agent's choices.

We are interested in how the agent's preferences – and subsequently his or her choices – depend on the reasons that motivate him or her. For the purposes of our theory, we think of a *reason* as a proposition which, if true

of an alternative and salient for an agent (in a sense that can be spelt out variously), may affect the agent's preferences for or against that alternative as compared to others. Examples of propositions that may serve as reasons are 'there is war', 'there is food available', 'the dish is poisonous', 'I am hungry', 'the power station has high CO_2 emissions', and so on. A reason which is currently salient for the agent in the appropriate sense – and whose effect on the agent's preferences is thereby operational – is called *motivating* for him or her. We write M to denote the agent's set of motivating reasons. This is a subset of the set of all possible reasons, which we call $\mathcal{R}$.[6]

To indicate the dependency of the agent's preferences on his or her set of motivating reasons, we append the subscript M to the symbol $\succsim$, interpreting $\succsim_M$ as the agent's preference order in the event that M is the agent's set of motivating reasons. Further, $\succ_M$ represents the corresponding strict preference, and $\sim_M$ the indifference relation.

3 Which reasons are motivating?

It is needless to say that, from a psychological perspective, not every possible reason will necessarily become motivating for a given agent. This depends very much on the agent's psychology and perhaps his or her environment. Moreover, from a normative perspective, not every reason might be deemed appropriate. Someone might be psychologically motivated, for example, by reasons which, morally speaking, we find deplorable. Conversely, a particular proposition might seem to be a compelling reason for or against some action from a third-person perspective – or perhaps from the perspective of some background normative theory – and yet it may fail to attain any motivational force for the agent. Since the aim of this paper is to develop

[6] The set $\mathcal{R}$ may coincide with the set of all propositions or be a proper subset of it, depending on how permissively or restrictively we interpret the notion of a reason. Our theory is completely general. It is consistent, for example, with the view that the set of reasons corresponds to what we might represent by certain atomic sentences in a suitable language, or with the view that it includes propositions corresponding to compound sentences, or even with the view that it satisfies certain closure properties (such as closure under conjunction or disjunction).

a *formal* theory of rational choice, rather than a theory of moral or normatively justified choice, we focus mainly on how an agent's motivating reasons *explain* his or her preferences and choices, rather than on whether there exist normative reasons *justifying* them and what those 'right' reasons are. Nonetheless, our theory also offers a useful framework for thinking about the latter question, as discussed later.

There are many possible accounts of when a reason becomes motivating for an agent, that is, when it is 'salient' for the agent so as to attain motivational force. For the purposes of this paper, we need not commit ourselves to a single such account. One possibility is that a reason becomes motivating for an agent as soon as he or she conceptualizes it abstractly – in the sense that, in his or her conceptualization of the world in the relevant context, the agent distinguishes between those alternatives of which the proposition in question is true and those of which it isn't. If our conceptualization of the world does not distinguish between those states of the world in which the number of grains of sand is even and those in which it is odd, for example, then the proposition that there exists an even number of grains of sand cannot be a motivating reason that affects our rational choices.[7]

Another possibility is that the abstract conceptualization of a given reason is not enough to make it motivating. Rather, a reason becomes motivating only when the agent qualitatively – and not merely abstractly – understands it. A policy maker, for example, may abstractly understand that different foreign policies can be distinguished from each other with respect to whether or not they make cheap oil available and whether or not they lead to war, but fail to understand qualitatively what a war involves and thus fail to be motivated by the latter reason. This second account of what makes a reason motivating, unlike the first one, requires that the dis-

[7] Generally, the agent's conceptualization of the world may be more coarse-grained than the one a suitably well-informed third-person observer is able to come up with. For instance, the agent him- or herself may only distinguish between certain non-singleton equivalence classes of alternatives in X rather than between individual alternatives. Consequently, only propositions expressible as unions of such equivalence classes may be conceptualized by the agent. According to our theory developed below, the agent would then be indifferent between alternatives in the same equivalence class.

tinction between abstract conceptualization and qualitative understanding can be meaningfully made – an issue which is partly philosophical and partly psychological; we flag it here as something that merits further investigation.

A third account draws on the concept of attentional salience as it is frequently used in psychology and behavioural economics. Among those propositions abstractly conceptualized by an agent and perhaps even 'understood' in some stronger, qualitative sense, only some are typically 'salient' for the agent, in that the agent actively focuses on them or uses them as a 'heuristic' in forming his or her preferences. Now the idea is that a reason becomes motivating for an agent if and only if he or she focuses on it actively or uses it as a preference formation heuristic. This account is particularly interesting in case the agent is boundedly rational, that is, unable or at least unlikely to give full and simultaneous attention to everything he or she conceptualizes or understands.[8] Whichever account of what makes a reason motivating we adopt, however, the basic idea that the agent's preferences depend on his or her set of motivating reasons is a very natural one.

4 The psychological states of an agent

From a third-person perspective, a full theory of an agent requires the ascription of an entire family of preference orders to that agent, consisting of one preference order $\succsim_M$ for each psychologically possible set of motivating reasons M. Different sets of motivating reasons thus correspond to different psychological states of the agent, and in each such state, the agent holds only one preference order. What the reference to an entire family of preference orders captures is the idea that the agent may have a disposition to change his or her preferences in certain ways when his or her set of motivating reasons changes. The policy maker in our earlier example may prefer an invasion of an oil-producing country to an investment in renewable resources

[8] A prominent account of rational choice based on heuristics has been developed by Gerd Gigerenzer, Peter M. Todd and the ABC Group, *Simple heuristics that make us smart* (New York: Oxford University Press, 1999). See also Gerd Gigerenzer and Reinhard Selten (eds.), *Bounded rationality: The adaptive toolbox* (Cambridge, MA: MIT Press, 2001).

if he or she is motivated only by whether the policy supports current consumption levels of cheap oil. However, this preference may change if he or she becomes motivated also by whether the policy leads to war. Of course, the agent him- or herself need not – and typically will not – be consciously aware of the entire family of preference orders ascribed to him or her by our theory.

In order to ascribe to the agent one preference order $\succsim_M$ for each possible set of motivating reasons M, we need to specify what the possible such sets are. In other words, we need to say something about what psychological states the agent can possibly be in. In the simplest case, every logically possible set of reasons, that is, every subset of $\mathcal{R}$, is a possible motivating set. However, we have already noted that it is a contingent psychological matter which reasons have the capacity to motivate a given agent, and under what conditions. As a result of such psychological constraints, not every subset of $\mathcal{R}$ needs to constitute a possible specification of M for the agent.[9] Moreover, there may be reasons that *can* motivate the agent, but never in conjunction with certain others. Perhaps some reasons crowd out other reasons, such as economic self-interest driven reasons versus charitable ones. Conversely, some types of reasons may motivate *only* in conjunction with specific others, and so on.

Thus, in the general case, the set of all possible sets of motivating reasons, which we call $\mathcal{M}$, may be smaller than the set of all subsets of $\mathcal{R}$. For the purpose of obtaining some general results, we make a regularity assumption about the possible sets of motivating reasons:

Regularity assumption. *The different possible sets of motivating reasons (that is, the elements of $\mathcal{M}$) form a lattice, that is:*

(i) if M_1 and M_2 are possible sets of motivating reasons, then so is $M_1 \cap M_2$ (that is, $\mathcal{M}$ is closed under intersection);

[9] It is not even strictly necessary that every reason in $\mathcal{R}$ occurs in at least one possible specification of M, although our theory allows us to define the set $\mathcal{R}$ tightly, so as to include only reasons that are at least sometimes motivational.

(ii) if M_1 and M_2 are possible sets of motivating reasons, then so is $M_1 \cup M_2$ (that is, $\mathcal{M}$ is closed under union).

In the baseline case in which all subsets of $\mathcal{R}$ are possible sets of motivating reasons, this regularity assumption is trivially satisfied. In fact, our formal analysis requires only something weaker than this assumption, but for expositional simplicity we set these details aside here.[10]

5 Two axioms on the relationship between reasons and preferences

We are now in a position to introduce our two central axioms on the relationship between an agent's set of motivating reasons and his or her preferences. The idea underlying both axioms is that an agent's preference for or against an alternative as compared with others is driven by the reasons that are true of the alternative *and* motivating for the agent. The first axiom concerns the case in which the exact same motivating reasons are true of a given pair of alternatives.

Axiom 1. *The agent is indifferent between any pair of alternatives of which the same motivating reasons are true. Formally, for any x and y in X and any M in $\mathcal{M}$, if $\{R \in M : R \text{ is true of } x\} = \{R \in M : R \text{ is true of } y\}$, then $x \sim_M y$.*

Axiom 1 is true almost by definition under two of the three illustrative interpretations of motivating reasons discussed above. Consider the 'conceptualization' interpretation. If the agent does not abstractly conceptualize

[10] As shown in the appendix, our first representation theorem below (theorem 1 and its subsequent extension 1*) uses only part (i) of the assumption and our second theorem (theorem 2 and its extension 2*) uses only a weakened variant of part (ii), which requires that if M_1 and M_2 are possible sets of motivating reasons, then so is some superset of $M_1 \cup M_2$. The latter condition is satisfied, for example, as soon as $\mathcal{R}$ is a possible motivating set. By suitably amending our axioms below, we can also say something about the case in which neither part of the assumption holds, but we do not discuss the details here.

any reasons that are not motivating for him or her, this means that he or she does not distinguish between alternatives that differ at most with respect to non-motivating reasons, and hence indifference between these alternatives is natural.[11] Similarly, consider the 'attentional salience' interpretation. If the agent gives no attention to any reasons that are not motivating for him or her or does not use them as a heuristic in forming his or her preferences, then it is only natural to expect that he or she will be indifferent between alternatives that are identical with respect to all motivating reasons. Under the remaining illustrative interpretation of motivating reasons, the 'qualitative understanding' interpretation, axiom 1 becomes a substantive – and we think interesting – psychological hypothesis. On this interpretation, axiom 1 says that an agent's preferences between alternatives are fully determined by those properties of the alternatives that the agent qualitatively understands, while any properties not understood in this manner make no difference.

Axiom 2 concerns the case in which the agent's set of motivating reasons grows, but none of the newly added reasons is true of a given pair of alternatives.

Axiom 2. *If additional reasons become motivating for the agent, but none of them are true of a given pair of alternatives, then the agent's preference over that pair remains unchanged. Formally, for any x and y in X and any M and M' in $\mathcal{M}$ with $M' \supseteq M$, if no R in $M' \backslash M$ is true of x or y, then $x \succsim_M y \Leftrightarrow x \succsim_{M'} y$.*

This implies, for example, that if the proposition 'Bordeaux wine is served at dinner' becomes a motivating reason for an agent, this does not affect his or her preference between two alternative dinner plans that do not involve any wine. The axiom is plausible under each of our three illustrative interpretations of motivating reasons, especially in light of the idea that an

[11] Indifference follows strictly under the following conditions: (i) the agent distinguishes only between possibly non-singleton equivalence classes of alternatives, so that his or her preferences over individual alternatives are induced by preferences over these equivalence classes, and (ii) his or her motivating reasons are precisely the propositions expressible as unions of such equivalence classes, as discussed in an earlier note.

agent's preferences between alternatives is driven by the reasons that are *both* motivating for the agent *and* true of those alternatives.

Apparent counterexamples to axiom 2 – that is, preference changes apparently driven by the addition of reasons that are not true of any of the alternatives in question – typically involve an under-specification of the reasons that are being added to the agent's motivating set. Consider, for instance, the following apparent counterexample. An agent currently prefers having dinner at McDonald's to dining at an organic vegetarian teetotalers' restaurant. But when he or she adopts the proposition 'Bordeaux wine is served at dinner' as a further motivating reason, this prompts in him or her a more sophisticated attitude towards food and thereby reverses his or her preference between McDonald's and the organic alternative. Since neither dinner option involves any wine, axiom 2 appears to be violated by this preference change. This appearance, however, rests on an under-specification of the additional motivating reasons that lead to the agent's new psychological state. Implicit in the example is the thought that, along with the proposition 'Bordeaux wine is served at dinner', a second proposition such as 'the food is sophisticated' has also gained motivational force, and it is the latter reason that is responsible for the preference change. This is entirely consistent with axiom 2, since one of the two dinner options offers sophisticated food.

6 A general representation of reason-based preferences

So what is the consequence of the two axioms we have introduced? Our first representation theorem shows that their satisfaction by an agent implies that the agent's family of preference orders across the different possible sets of motivating reasons can be represented in terms of a single underlying binary relation over consistent combinations of reasons. Formally, call a set of reasons S (a subset of $\mathcal{R}$) *consistent* if there exists at least one alternative x in X of which all the reasons in S are true.

Theorem 1. *The agent's preference orders $\succsim_M$ across the different possible sets of motivating reasons M in $\mathcal{M}$ satisfy axioms 1 and 2 if and only if there exists a binary relation $\geq$ over all consistent sets of reasons such that, for each M in $\mathcal{M}$,*

$$x \succsim_M y \;\Leftrightarrow\; \{R \in M : R \text{ is true of } x\} \geq \{R \in M : R \text{ is true of } y\} \text{ for all } x,y \text{ in } X.$$

Under the theorem's conditions, the agent's entire family of preference orders can thus be generated by a single binary relation $\geq$ over consistent reason sets, which we accordingly call a *generating relation*. For any given set of motivating reasons, the agent prefers an alternative x to another alternative y just in case the generating relation ranks the set of reasons that are true of x and motivating above the set of reasons that are true of y and motivating.

Before we turn to the interpretation of this relation, it is useful to give an example. Consider a very simple case in which there are only four possible alternatives over which the agent has preferences:

$$\begin{aligned} ht &: \text{I am healthy, and you are tired.} \\ \neg ht &: \text{I am not healthy, and you are tired.} \\ h\neg t &: \text{I am healthy, and you are not tired.} \\ \neg h\neg t &: \text{I am not healthy, and you are not tired.} \end{aligned}$$

Suppose, further, that the set $\mathcal{R}$ of possible reasons contains only two reasons, namely

$$\begin{aligned} H &= \{ht, h\neg t\} \text{ (I am healthy), and} \\ T &= \{ht, \neg ht\} \text{ (you are tired),} \end{aligned}$$

but any combination of them can be motivating, that is, the set $\mathcal{M}$ consists of all subsets of $\mathcal{R}$. Now imagine that the agent's preferences across the different possible sets of motivating reasons are as follows:

$$\begin{aligned} M = \{H,T\} &\Rightarrow h\neg t \succ_M ht \succ_M \neg h\neg t \succ_M \neg ht; \\ M = \{H\} &\Rightarrow h\neg t \sim_M ht \succ_M \neg h\neg t \sim_M \neg ht; \\ M = \{T\} &\Rightarrow h\neg t \sim_M \neg h\neg t \succ_M ht \sim_M \neg ht; \\ M = \varnothing &\Rightarrow h\neg t \sim_M ht \sim_M \neg h\neg t \sim_M \neg ht. \end{aligned}$$

One can verify that these preferences do indeed satisfy axioms 1 and 2, so that our theorem applies. What, then, does the underlying generating relation look like? It is easy to check that the agent's preference orders across the different possible sets of motivating reasons just displayed can be generated by a single binary relation $\geq$ over consistent reason sets that satisfies

$$\{H\} > \{H, T\} > \varnothing > \{T\},$$

where $>$ denotes the strict relation induced by $\geq$. (When applicable, we further write $\equiv$ for the symmetrical relation induced by $\geq$.) Informally, the reason set $\{H\}$ is ranked first, the set $\{H, T\}$ is ranked second, the empty set is ranked third, and the set $\{T\}$ is ranked last, which captures a particular way of 'weighing' of these reason sets relative to each other.[12]

But how can the agent's generating relation over consistent reason sets be interpreted? We can distinguish between at least two broadly different kinds of interpretations. According to the first kind of interpretation, which may be described as 'cognitivist', a generating relation encodes a particular set of propositions about the relative 'goodness' of different sets of reasons. In particular, $S_1 \geq S_2$ is taken to mean that S_1 is a (weakly) better set of reasons than S_2. Depending on the precise variant of this interpretation, these propositions may capture either agent-independent or agent-dependent *facts* about the goodness of different reason sets, or some implicit goodness *judgments* attributable to the agent.

According to the second kind of interpretation of a generating relation, which we may call 'non-cognitivist', a generating relation encodes the agent's dispositions to prefer certain sets of reasons over others when the reasons contained in them are motivating. Here, $S_1 \geq S_2$ is taken to mean that the reason set S_1 is (weakly) preferred to the reason set S_2 whenever all reasons in S_1 and all those in S_2 are motivating. Again, different variants of this interpretation are conceivable, depending on what precisely is understood by a preference over reason sets.

Regardless of the interpretation we ultimately adopt, however, our theorem shows that when the relationship between an agent's set of motivating

[12] The special case of 'additive' weighing is considered in a separate section below.

reasons and his or her preferences is governed by our two axioms, those preferences can be parsimoniously represented in terms of a single underlying generating relation over consistent reason sets. Furthermore, this generating relation is essentially unique. That is to say, it is unique on those pairs of consistent reasons sets that are actually needed to generate the agent's preference orders across different sets of motivating reasons.[13] In short, our theorem delivers a simple representation of what is by itself a rich structure, namely the agent's family of preference orders across all possible sets of motivating reasons.

7 Is the generating relation transitive?

Although we have considered different possible interpretations of the agent's generating relation, we have not said anything yet about its formal properties. Most importantly, is the generating relation actually an order over the consistent reason sets? In other words, is it a complete and transitive binary relation? Completeness turns out to be not much of a problem since the generating relation can always be defined so as to (weakly) rank all pairs of consistent reason sets. Surprisingly, however, the conditions introduced so far do not guarantee that the generating relation will always be transitive, despite the fact that all the actual preference orders generated by it are transitive. So how can an intransitivity in the generating relation occur, and when is it ruled out?

To address these questions, it is helpful to begin with an example. Consider an agent who forms preferences over three types of cars:

$fb\neg e$: a car that is fast, big, but not environmentally friendly;

$f\neg be$: a car that is fast, not big, but environmentally friendly;

$\neg fbe$: a car that is not fast, but big and environmentally friendly.

[13] These are all the pairs of reason sets expressible as $\{R \in M : R \text{ is true of } x\}$ and $\{R \in M : R \text{ is true of } y\}$ for some x, y in X and M in $\mathcal{M}$. The generating relation is underdetermined only with respect to those pairs of reason sets that cannot be instantiated as true of some actual alternatives in X and motivating, and such pairs do not really matter from the perspective of the agent's rational choices.

Thus any car available on the market has precisely two out of the three characteristics: fast (f), big (b), and environmentally friendly (e). Suppose, further, that a car's having any one of these characteristics can serve as a reason for or against preferring it, that is, the different possible reasons in $\mathcal{R}$ are

$$\begin{aligned} F &= \{fb\neg e, f\neg be\} \text{ (the car is fast)}, \\ B &= \{fb\neg e, \neg fbe\} \text{ (the car is big)}, \\ E &= \{f\neg be, \neg fbe\} \text{ (the car is environmentally friendly)}. \end{aligned}$$

Moreover, we assume that any combination of these reasons can be motivating, that is, $\mathcal{M}$ contains all subsets of $\mathcal{R}$. Now it is entirely conceivable that the agent's preferences across the different possible sets of motivating reasons are the following:

$$\begin{aligned} M = \{F, B, E\} &\Rightarrow fb\neg e \sim_M f\neg be \sim_M \neg fbe, \\ M = \{F, B\} &\Rightarrow fb\neg e \succ_M f\neg be \succ_M \neg fbe, \\ M = \{B, E\} &\Rightarrow \neg fbe \succ_M fb\neg e \succ_M f\neg be, \\ M = \{F, E\} &\Rightarrow f\neg be \succ_M \neg fbe \succ_M fb\neg e, \\ M = \{F\} &\Rightarrow fb\neg e \sim_M f\neg be \succ_M \neg fbe, \\ M = \{B\} &\Rightarrow fb\neg e \sim_M \neg fbe \succ_M f\neg be, \\ M = \{E\} &\Rightarrow f\neg be \sim_M \neg fbe \succ_M fb\neg e, \\ M = \varnothing &\Rightarrow fb\neg e \sim_M f\neg be \sim_M \neg fbe. \end{aligned}$$

One can check without too much difficulty that these preferences satisfy axioms 1 and 2,[14] and so, by theorem 1, they are representable in terms of an underlying generating relation $\geq$ over consistent reason sets. But what is this generating relation? In order to be able to generate the agent's preferences just displayed, it must have all of the following properties:

[14] This is very straightforward in the case of axiom 1. To see that axiom 2 is satisfied, notice that the structure of the example implies that whenever x, y in X, M in $\mathcal{M}$ and R in $\mathcal{R}\backslash M$ are such that R is true of neither x nor y, then x and y must be the same alternative (that is, $x = y$), so that $x \sim_M y$ and $x \sim_{M\cup\{R\}} y$.

$$\{F, B\} \equiv \{B, E\} \equiv \{F, E\},$$

$$\{F, B\} > \{F\} > \{B\},$$
$$\{B, E\} > \{B\} > \{E\},$$
$$\{F, E\} > \{E\} > \{F\},$$

$$\{F\} > \varnothing,$$
$$\{B\} > \varnothing,$$
$$\{E\} > \varnothing.$$

From the second, third and fourth lines, it follows immediately that the generating relation violates transitivity, since $\{F\} > \{B\}$, $\{B\} > \{E\}$, and yet $\{E\} > \{F\}$. Of course, an intransitive generating relation is harder to interpret than a transitive one, particularly if one chooses to adopt a cognitivist interpretation. The lesson to learn, however, is that the conditions used in our first representation theorem are simply not enough to rule out an intransitive generating relation, though they are fully compatible with an agent's having a transitive such relation.[15]

What is the source of the intransitivity in the agent's generating relation in our example? Imagine that, contrary to the assumptions made, there existed additional cars of which precisely one or none of the three reasons considered is true: one car that is only fast, one that is only big, one that is only environmentally friendly, and one without any of these properties. Then the agent's preference order over the different possible cars when motivated by all three reasons would constrain his or her generating relation to rank all consistent reason sets, including the three sets $\{F\}$, $\{B\}$ and $\{E\}$,

[15] Notice that an intransitivity in an agent's generating relation does *not* give rise to an intransitivity in the agent's actual preferences, so long as – and this is an important qualification – the agent's preferences are stably defined relative to a fixed set of motivating reasons. If, however, the agent's consideration of different pairs of alternatives somehow endogenously led to a shift in the agent's set of motivating reasons, then an intransitivity might well surface. This would happen, for example, if the comparison of $f\neg be$ and $\neg fbe$ made the set $\{F, B\}$ motivating (leading to a preference for $f\neg be$ over $\neg fbe$), the comparison of $\neg fbe$ and $fb\neg e$ made the set $\{F, E\}$ motivating (leading to a preference for $\neg fbe$ over $fb\neg e$), and the comparison of $fb\neg e$ and $f\neg be$ made the set $\{B, E\}$ motivating (leading to a preference for $fb\neg e$ over $f\neg be$). We return to this issue briefly in the last section of this paper.

transitively. The intransitivity identified in our example would disappear. The counterfactual stipulation just made would give the agent a kind of 'Olympian perspective' from which he or she would be able to consider one alternative corresponding to each consistent reason set, which instantiates all and only the reasons in it, and to rank all these sets transitively. Generalizing from this observation, we can conjecture that an intransitivity in the agent's generating relation can occur precisely if this Olympian perspective is not available.

Our second representation theorem confirms this conjecture. Call the set $\mathcal{R}$ of reasons *weakly independent* if, for every consistent subset S of $\mathcal{R}$, it is possible for the reasons in S to be true while all the reasons in $\mathcal{R}\backslash S$ are false – or equivalently, there exists at least one alternative x in X of which all the reasons in S and no others are true. To illustrate, weak independence of the reasons in $\mathcal{R}$ is violated in our example of the three cars: Even though each of the three reason sets $\{F\}$, $\{B\}$ and $\{E\}$ (as well as the empty set) is consistent, there do not exist cars instantiating each such reason set. By contrast, in the augmented example in which cars instantiating each such set are stipulated to exist, weak independence is satisfied.

Theorem 2. *Suppose the set of reasons $\mathcal{R}$ is weakly independent. Then the agent's preference orders $\succsim_M$ across the different possible sets of motivating reasons M in $\mathcal{M}$ satisfy axioms 1 and 2 if and only if there exists an order $\geq$ (a complete and transitive binary relation) over all consistent sets of reasons such that, for each M in $\mathcal{M}$,*

$$x \succsim_M y \;\Leftrightarrow\; \{R \in M : R \text{ is true of } x\} \geq \{R \in M : R \text{ is true of } y\} \text{ for all } x, y \text{ in } X.$$

The theorem confirms that in our example the lack of weak independence of the set of possible reasons is indeed to blame for the unavailability of the Olympian perspective needed to ensure a transitive generating relation. Conversely, the satisfaction of weak independence is enough to guarantee the transitivity of the generating relation. As in our earlier representation theorem, the generating relation – now a *generating order* – is essentially unique.

8 Additive weighing of reasons

It is worth drawing attention to one important special case of an agent who is rational in the sense of our theory, and to whom our representation results therefore apply. Imagine an agent whose reason-based preference formation works as follows. The agent implicitly assigns a particular 'value' or 'weight' to each of the possible reasons in $\mathcal{R}$. Some reasons get assigned a positive value, others a negative one. For example, the reason 'there is peace' will presumably have a positive value, while the reason 'there is not enough food available' or 'I am hungry' will have a negative one. Now the agent prefers one alternative to another just in case the sum-total of the values or weights assigned to the reasons that are true of the first alternative and motivating exceeds the same sum-total for the second alternative. In each case, the sum-total encompasses both positively and negatively valued reasons.

More formally, this process can be described as follows. The values or weights assigned to the reasons in $\mathcal{R}$ are represented by a function v from the set of reasons $\mathcal{R}$ into the real numbers, which assigns to each reason R in $\mathcal{R}$ a particular (positive or negative) value $v(R)$. Of course, this function may differ from agent to agent. For each set of motivating reasons M in $\mathcal{M}$, the agent's preference order $\succsim_M$ is now given as follows:

$$x \succsim_M y \Leftrightarrow \sum_{R \in M : R \text{ is true of } x} v(R) \geq \sum_{R \in M : R \text{ is true of } y} v(R) \quad \text{for all } x, y \text{ in } X.$$

Our two axioms are clearly satisfied here, and the agent's generating relation over consistent reasons sets can easily be derived from the values or weights assigned to each of the different reasons contained in them. Specifically, one set of reasons is ranked over another by the agent's generating relation just in case the sum-total of values or weights assigned to the reasons in the first set exceeds that for the second, or formally:

$$S_1 \geq S_2 \Leftrightarrow \sum_{R \in S_1} v(R) \geq \sum_{R \in S_2} v(R) \quad \text{for any consistent } S_1, S_2 \subseteq \mathcal{R}.$$

To illustrate, recall our earlier example of the agent holding preferences over the four alternatives corresponding to the different possible truth-value

combinations of the propositions 'I am healthy' and 'you are tired'. In that example, the agent's preferences and underlying generating relation can be represented in terms of the assignment of suitable values or weights to individual reasons. It is a straightforward exercise to check that we obtain a correct representation of the given preferences by assigning, for example, a weight of 2 to the reason 'I am healthy' (H) and a weight of -1 to the reason 'you are tired' (T).

To be sure, an agent whose reason-based preference formation works like this is only a special case of a rational agent as described by our theory, since, for the agent in the current example, reasons have an 'additive separability' property that they need not have in general: In this special case, the value or weight the agent assigns to any motivating reason is independent of what other reasons are motivating for him or her. In the general case described by our theory, there is no such restriction.

Nonetheless, the special case we have flagged is of interest since the additive balancing of reasons that goes on here captures what in philosophical discussions is often described as 'weighing of reasons', particularly the weighing of *pro tanto* reasons for and against some object of choice.[16] Indeed, it is only in the context of separability that any given reason can unambiguously be said to 'count in favour of' or 'against' the alternatives of which it is true. Without separability, the question of whether a reason counts for or against those alternatives depends entirely on which other motivating reasons are present.

9 Explanatory versus justificatory reasons

It is appropriate at this point to take a step back and to revisit the interpretation of our reason-based approach to the theory of rational choice. As we have pointed out, our theory is intended to be a formal theory of rational choice, not a theory of moral or normatively justified choice, and hence our focus is on how an agent's motivating reasons *explain* his or her preferences and choices, rather than on whether there exist any normative reasons *justi-*

[16] See, for example, the discussion of *pro tanto* reasons in Broome, 'Reasons' (op. cit.).

fying them. Despite this focus, our theory also provides a useful conceptual framework for thinking about such more normative concerns.

To show how our theory can speak to those concerns, let us begin with the distinction between *explanatory* and *justificatory* reasons.[17] The former are the reasons that *explain* why an agent has *actually* made the choices he or she has made; the latter are the reasons, if any, that would *justify* those choices from the perspective of some normative background theory, say a moral one. Recall our example of a policy maker who is deciding which foreign policy to support. It may happen, as we have noted, that he or she supports the invasion of an oil-producing country because this promises to make cheap oil available. But since it is hard to think of any mainstream theory of just war that would deem an invasion justified on those grounds, the present case illustrates how explanatory and justificatory reasons can come apart. While the prospective availability of cheap oil is certainly an *explanatory* reason for the policy maker's decision – it explains his or her preference – it is by no means a *justificatory* one: It does not normatively justify that preference.

From a purely social-scientific perspective, we are primarily interested in explaining why agents make the choices they make, and sometimes in predicting those choices. From a philosophical perspective, however, we would ideally like to do more than that: We would also like to assess whether an agent's choices are justified – or at least whether they are justifiable – and whether the agent has made them for the right reasons. Implicit in the aim to subject the agent's choices to such normative assessment is the idea that there are not only the generating relation that *actually* governs the agent's preference formation and the set of reasons that *actually* motivates him or her, but that there are also some independent criteria for determining what kind of generating relation is normatively permissible and what kinds of motivating reasons are normatively required if the agent is to act 'for the right reasons'.

[17] On this distinction, see also Lenman, 'Reasons for Action: Justification vs. Explanation' (op. cit.).

It is compatible with our theory to assume that only some of the possible generating relations over consistent reason sets are permissible, and similarly that, in any given context, only particular sets of motivating reasons are appropriate.[18] We would then be able to distinguish between choices that can be explained but not justified in a reason-based manner, and choices that can also be so justified. We have seen that, as soon as an agent's preferences satisfy axioms 1 and 2, we can give a reason-based *explanation* of those preferences, using our representation theorems. However, whether an agent's preferences, say between x and y, are also *justified* depends on whether the agent's actual generating relation (or at least that portion of the relation driving the preferences in question) is normatively permissible *and* whether the agent's actual set of motivating reasons is a normatively appropriate one. We are also able to describe those cases in which an agent's preferences are at least *justifiable*, even when the more stringent conditions of actual justification may be violated. An agent's preferences are *justifiable* if there *exist* a normatively permissible generating relation and a normatively appropriate set of motivating reasons that would give rise to those preferences, independently of whether or not this is the actual way in which the agent has arrived at them.

Thus our theory not only allows us to distinguish between explicable and justifiable choices, and between choices made for the right and the wrong reasons, but it also gives expression to the less commonly recognized possibility that an agent is motivated by the right reasons but governed by the wrong underlying generating relation, or that the agent's being motivated by the wrong reasons goes along with his or her having the right generating relation.

While the distinctions between explanatory and justificatory reasons and between acting for the right and the wrong reasons are familiar from the existing philosophical literature, the role played by the agent's generating relation, and the additional complexities that open up once we subject this to a normative assessment as well, are usually not made explicit in the

[18] On a very strong version of this assumption, there exists only one right generating relation and one right set of motivating reasons for each context.

literature. It should therefore be evident that our proposed theory offers useful conceptual resources for addressing those under-researched issues.

10 Deliberation and disagreements

Just as our theory allows us to discuss reasons from both explanatory and justificatory perspectives, so it can also shed light on the role played by reasons in an agent's rational deliberation about his or her preferences and on the relevance of reasons when there are disagreements between different agents' preferences. As we have noted, standard rational choice theory takes preferences over possible worlds or fully specified outcomes – the *alternatives* in our theory – to be fundamental and unchangeable and thus cannot explain how deliberation, either within a single individual or in a group, could possibly lead to any revisions of those preferences. It is assumed that preference changes can only ever take place at the level of derived preferences over non-fundamental prospects and must stem from changes in the agent's beliefs about which outcomes are likely to result from those prospects. By implication, when individuals engaged in collective deliberation have different preferences, all we can do is to resolve any informational differences between them and, if this does not help to reach agreement (because the disagreement was not due to different information), to aggregate their conflicting fundamental preferences into overall collective preferences. This, however, does not resolve the individual-level disagreement; at best, it generates some collective compromise. Furthermore, the process is notoriously vulnerable to the paradoxes and impossibility results of aggregation familiar from social choice theory in the tradition of Condorcet and Arrow.[19]

This standard picture fails to account for the possibility that an agent's preferences may change as a result of a change in his or her motivational state, which, in turn, may be prompted by various experiences and especially by individual or collective deliberation. We can think, for example, of a capitalist businessman who, after surviving a plane crash, consciously forms a preference for a life devoted to charity over a life driven by income

[19] Kenneth Arrow, *Social Choice and Individual Values* (New York: Wiley, 1951/1963).

maximization, or a workaholic who, after recovering from an illness, consciously abandons his or her work-oriented preferences.[20] Similarly, group deliberation may change participants' assessment of fundamental alternatives, for instance by making previously overlooked aspects of those alternatives salient to them.[21] Arguably, these agents have not merely learnt new information – although some of their beliefs may have changed along the way – but new reasons, such as other-regarding or non-economic reasons, have gained motivational force for them.

Our reason-based theory of rational choice allows us to capture those phenomena. It allows us to distinguish between information-based and reason-based deliberation, and to acknowledge mixtures of the two. Information-based deliberation, as recognized by standard rational choice theory, takes place whenever an agent rationally revises his or her beliefs in response to new information or evidence, which may affect the agent's derived preferences over non-fundamental prospects, but not his or her preferences over fundamental alternatives. Reason-based deliberation, on the other hand, takes place when the agent rationally revises his or her preferences at the fundamental level in response to changes in his or her set of motivating reasons.[22]

On this picture, an agent can enter into a conscious deliberative process with the aim of identifying which reasons to use in forming his or her pref-

[20] For a more extensive discussion of the phenomenon of preference change, see Franz Dietrich and Christian List, 'A Model of Non-Informational Preference Change', working paper, London School of Economics (2009).

[21] For related discussions, see David Miller, 'Deliberative Democracy and Social Choice', *Political Studies* 40 (special issue) (1992), pp. 54-67; Jack Knight and James Johnson, 'Aggregation and Deliberation: On the Possibility of Democratic Legitimacy', *Political Theory* 22 (1994), pp. 277-296; John S. Dryzek and Christian List, 'Social Choice Theory and Deliberative Democracy: A Reconciliation', *British Journal of Political Science* 33 (1) (2003), pp. 1-28; and Christian List, Robert Luskin, James Fishkin and Iain McLean, 'Deliberation, Single-Peakedness, and the Possibility of Meaningful Democracy: Evidence from Deliberative Polls', working paper, London School of Economics and Stanford Center for Deliberative Democracy (2000/2007).

[22] For an earlier taxonomy of informational, argumentative, reflective, and social aspects of deliberation, see Dryzek and List, 'Social Choice Theory and Deliberative Democracy: A Reconciliation' (op. cit.).

erences over a given set of alternatives. Although someone not engaged in explicit deliberation may sometimes simply find him- or herself being motivated by some reasons rather than others, we need not assume that one's set of motivating reasons is always outside one's control. Deliberation can lead an agent explicitly to reflect on what reasons matter, and thereby to exercise some influence over which reasons come to motivate him or her.

This is entirely consistent with each of our three identified accounts of when a reason becomes motivating. Consider the 'conceptualization' account, according to which an agent gets motivated by a given reason as soon as he or she conceptualizes it abstractly. Consistently with this account, deliberation may affect an agent's set of motivating reasons through refining his or her conceptual abilities, that is, by leading him or her to learn how to distinguish between alternatives of which particular reasons are true and alternatives of which they are not. Similarly, consider the 'qualitative understanding' account, according to which a reason becomes motivating when the agent qualititatively – and not merely abstractly – understands it. Although developing this idea would require further elaboration, it is plausible that deliberation, at least when construed broadly, is not restricted to the exchange of information or to abstract conceptual reasoning, but that it can also make an agent imagine various scenarios vividly and thereby enhance his or her qualitative understanding of some of the propositions true of those scenarios.[23] Think, for example, of how a startling personal report of someone's experience – say, the experience of war – can evoke in the listener a qualitative sense of what it might have been like to go through that experience oneself. Finally, in the case of the 'attentional salience' account of motivation, our claim that deliberation can affect an agent's set of motivating reasons should be fairly evident. If reasons become motivating whenever they are sufficiently salient for the agent, then obviously any activity, such as deliberation, that involves giving careful attention to various propositions

[23] For suggestions along these lines, see, for example, Iris Marion Young, *Intersecting voices: dilemmas of gender, political philosophy, and policy* (Princeton: Princeton University Press, 1997), and Robert E. Goodin, 'Democratic Deliberation Within', *Philosophy and Public Affairs* 29 (1) (2000), pp. 81-109.

can thereby push some of those propositions above the required threshold of motivational salience and deemphasize others. It is less clear, but still an open question, whether deliberation may affect not only the agent's motivating reasons, as we have argued, but also his or her underlying generating relation itself. We need not take a view on this issue here, except to mention it for further investigation.

The present observations point to a more nuanced perspective on the case of disagreements between different agents' preferences. As we have seen, when agents disagree in their preferences even after exchanging all relevant information, standard rational choice theory offers no further resources for resolving that disagreement – except perhaps to apply some method of 'brute' preference aggregation for arriving at overall collective preferences in the face of such disagreement. Our theory, by contrast, allows us to pinpoint precisely whether the disagreement stems from differences in the agents' sets of motivating reasons or from differences in their underlying generating relations (or both).

If it stems from differences in the agents' sets of motivating reasons, then it falls under the scope of deliberation in the broadened sense of our theory. The agents can deliberate about which reasons are the right ones to use in forming their preferences over the alternatives in question, and if they reach agreement on the right reasons, their disagreement will have been resolved. But even if they cannot fully agree on the appropriate reasons, their disagreement will at least have been made more tractable. The source of their disagreement will have been identified, which means that the disagreement will no longer have to be attributed to a brute difference in tastes. Much of the recent debate on the idea of 'public reason' can be understood along these lines: The aim is to come up with criteria for determining which reasons are publicly acceptable – that is, which reasons can be invoked to support one's preferences in public deliberation – and which are not.

If the agents' disagreement stems from differences in their underlying generating relations, on the other hand, the situation is more complicated. We have left it open to what extent deliberation can affect one's generating relation, but it should be noted that many debates in moral philosophy can

be understood as subjecting that relation to normative assessment as well. Such assessment takes place whenever the relative importance or weight of different reasons is being discussed.

Combining these observations, our theory further allows us to suggest a new approach to the aggregation of preferences: reason-based preference aggregation. Here each agent's preference order would be treated not as a fundamental and unchangeable input to the aggregation, but as being derived from two more fundamental inputs: a set of motivating reasons and an underlying generating relation. By making explicit and disentangling these two determinants behind any preference order, the informational basis for the aggregation would be enriched, which in turn might allow us to find more compelling methods of aggregation. Although this proposal clearly needs to be elaborated further, many theorists of 'public reason' may be attracted to the idea of aggregating preferences in a way that is sensitive to the reasons behind those preferences.

11 Preferences over uncertain prospects

A satisfactory theory of rational choice must be able to say something not only about an agent's preferences over possible worlds or fully specified outcomes – that is, over the alternatives we have focused on so far – but also about his or her preferences over uncertain prospects. In the real world, an agent's objects of choice are typically different possible actions, which correspond to different such prospects. Each action usually has a number of possible outcomes, and the agent is at most able to assign probabilities to them. These probabilities normally represent the agent's beliefs about the likelihood of those outcomes, but they could also have an objective interpretation. The aim of this section is to show that our theory can be extended so as to capture preferences over prospects in full generality. Less technically inclined readers may skip the material in this section without losing the overall thread of our argument.

Formally, we define a *prospect* as a probability distribution over the alternatives in X, that is, a function P from the set X of alternatives into

the interval $[0, 1]$ whose sum-total across all alternatives is 1.[24] We write $\mathcal{X}$ to denote the set of all such prospects. We now assume that an agent's preference order $\succsim_M$, for any possible set of motivating reasons M, is defined not just over the alternatives in X, but over all prospects in $\mathcal{X}$. Moreover, we assume that, for any fixed set of motivating reasons M, the agent's preferences are 'classical', in the sense that $\succsim_M$ ranks prospects according to the expectation of some 'utility' function from the set X of alternatives into the real numbers.

Again, we impose two axioms on the agent's preferences. The first axiom is identical in meaning to our earlier axiom 1. This is because it quantifies over all sure prospects, where a *sure prospect* is simply a prospect that assigns probability 1 to a single alternative x in X and probability 0 to all others; it can thus be identified with that alternative x.

Axiom 1*. *The agent is indifferent between any pair of sure prospects of which the same motivating reasons are true. Formally, for any sure prospects x and y in $\mathcal{X}$ and any M in $\mathcal{M}$, if $\{R \in M : R$ is true of $x\} = \{R \in M : R$ is true of $y\}$, then $x \sim_M y$.*

The second axiom quantifies over *all* prospects, not merely the sure ones, but otherwise matches our earlier axiom 2. To state this axiom, we define the probability that a prospect assigns to a given reason as the sum of the probabilities it assigns to the alternatives of which that reason is true.[25]

Axiom 2*. *If additional reasons become motivating for the agent, but all of them are assigned zero probability by a given pair of prospects, then the agent's preference over that pair remains unchanged. Formally, for any prospects P and Q in $\mathcal{X}$ and any M and M' in $\mathcal{M}$ with $M' \supseteq M$, if all R in $M' \backslash M$ receive zero probability under P and Q, then $P \succsim_M Q \Leftrightarrow P \succsim_{M'} Q$.*

When the agent's preferences satisfy axioms 1* and 2*, two representation theorems hold, which are direct analogues of our earlier two theorems.

[24] We require specifically that P has finite support, i.e., there are only finitely many x in X for which $P(x) > 0$.

[25] Formally, for any R in $\mathcal{R}$, $P(R) := \sum_{x \in X : R \text{ is true of } x} P(x)$.

Let us begin with the analogue of theorem 1, which provides a representation of the agent's preferences in terms of a single generating relation, this time defined not over consistent reason sets themselves, but more generally over probability distributions over consistent reason sets.[26] Recall that, in theorem 1, the preference between two alternatives x and y was determined by comparing the two reason sets $\{R \in M{:}R \text{ is true of } x\}$ and $\{R \in M{:}R \text{ is true of } y\}$. In the extension to the case of uncertainty, the preference between two prospects P and Q is determined by comparing two induced probability distributions over consistent reason sets, which we call P_M and Q_M. These can be seen as the probabilistic generalizations of the two reason sets compared in theorem 1. Specifically, P_M and Q_M assign to each consistent reason set S the total probability (according to P and Q, respectively) of those alternatives x in X for which S is the set of true motivating reasons.[27] In the special case in which P and Q are sure prospects – and thus identifiable with some alternatives x and y – the induced distributions P_M and Q_M assign probability 1 to $\{R \in M{:}R \text{ is true of } x\}$ and $\{R \in M{:}R \text{ is true of } y\}$, respectively. Now the analogue of theorem 1 can be stated as follows:

Theorem 1*. *The agent's preference orders $\succsim_M$ across the different possible sets of motivating reasons M in $\mathcal{M}$ satisfy axioms 1* and 2* if and only if there exists a binary relation $\geq$ over all probability distributions over consistent reason sets such that, for each M in $\mathcal{M}$,*

$$P \succsim_M Q \;\Leftrightarrow\; P_M \geq Q_M \text{ for all } P,Q \text{ in } \mathcal{X},$$

where P_M and Q_M are the induced probability distributions just defined.

[26] A probability distribution over consistent reason sets is a function from the set of all consistent reason sets to the interval $[0, 1]$ which sums to 1, where, as before, only finitely many consistent reason sets are assigned non-zero probability.

[27] In probability theory, P_M and Q_M are known as the projections of P and Q under the function that maps each alternative x in X to the consistent reason set $\{R \in M : R \text{ is true of } x\}$. Formally, for each consistent reason set S,

$$P_M(S) = \sum_{\substack{x \in X: \\ \{R \in M:R \text{ is true of } x\} = S}} P(x) \quad \text{and} \quad Q_M(S) = \sum_{\substack{x \in X: \\ \{R \in M:R \text{ is true of } x\} = S}} Q(x).$$

While theorem 1* shows that the satisfaction of axioms 1* and 2* is enough to ensure that the agent's preferences can be represented in terms of a single underlying generating relation, we can obtain a stronger representation when the set of reasons $\mathcal{R}$ is weakly independent, as before. Recall that weak independence of $\mathcal{R}$ means that, for every consistent subset S of $\mathcal{R}$, it is possible for the reasons in S to be true while all the reasons in $\mathcal{R}\backslash S$ are false. The following can be seen as the analogue of our earlier theorem 2, but, unlike theorem 2, it yields a representation of the agent's preferences in terms of a value function v over consistent reason sets, not just in terms of a generating order $\geq$ over them. Although the value function v always implies a generating order $\geq$, which ranks probability distributions over consistent reason sets according to their expected value under v, it contains more information than that order.

Theorem 2*. *Suppose the set of reasons $\mathcal{R}$ is weakly independent. Then the agent's preference orders $\succsim_M$ across the different possible sets of motivating reasons M in $\mathcal{M}$ satisfy axioms 1* and 2* if and only if there exists an underlying value function v from the set of consistent reason sets into the real numbers such that, for each M in $\mathcal{M}$, $\succsim_M$ ranks prospects according to the expected value of the set of true motivating reasons.*[28]

Thus the extension of our theory to the case of preferences over general prospects adds some further structure to the second representation theorem as compared with its earlier counterpart. In theorem 2 above, as noted, weak independence of the set of reasons $\mathcal{R}$ ensured an *ordinal* representation of the agent's preferences in terms of a single generating order over consistent reason sets. By contrast, theorem 2* yields a *cardinal* representation of those preferences in terms of a single value function v over consistent reason sets. The agent's preferences over prospects are then determined by the expected value of the set of true motivating reasons under those prospects.

[28] This expected value is the expectation of the induced utility function u_M mapping each alternative x in X to $v(\{R \in M : R \text{ is true of } x\})$. The value function v is unique up to positive affine transformations on the subdomain of those consistent reason sets needed to generate the agent's preferences. The functions u_M (for all M in $\mathcal{M}$) are unique up to the same transformations on the full domain.

We can redescribe this representation in another way. For each possible set of motivating reasons M, the value function v can be interpreted to induce a utility function u_M on the set of alternatives X. This utility function assigns to each alternative the value of the set of reasons that are true of that alternative and motivating, formally

$$u_M(x) = v(\{R \in M : R \text{ is true of } x\}) \quad \text{for each } x \text{ in } X.$$

The agent's preference order $\succsim_M$ over prospects is then determined by the expectation of that utility function for the given prospects, formally

$$P \succsim_M Q \Leftrightarrow \sum_{x \in X} P(x)u_M(x) \geq \sum_{x \in X} Q(x)u_M(x) \quad \textit{for all } P,Q \textit{ in } \mathcal{X}.$$

The present results demonstrate not only that our theory is able to represent preferences over uncertain prospects as much as it can represent preferences over sure ones, but also that it properly generalizes standard rational choice theory. Indeed, standard rational choice theory emerges as the special case of our theory in which the set M of motivating reasons is assumed to be fixed and sufficiently large to impose no restrictions on the assignment of utilities to individual alternatives.

12 From reason-based preferences to reason-based choices

We have given a detailed account of the relationship between an agent's set of motivating reasons and his or her preferences. In order to complete our theory of rational choice, we finally need to make explicit the way in which the agent's preferences affect his or her choices. To do so, we use a familiar idea from decision theory: the idea that any preference order induces a corresponding choice function. While a preference order represents certain *intentional attitudes* an agent holds towards a given set of alternatives, it does not explicitly tell us what the agent's *choice behaviour* would be when faced with a choice between some of those alternatives. A *choice function* provides a formal representation of that choice behaviour. Formally, it is a

function, denoted C, which assigns to each set of alternatives that the agent may be confronted with (that is, to each non-empty subset Y of X) a set of one or more chosen alternatives (that is, a non-empty subset of Y).

For example, suppose an agent is faced with a choice between different fruits, such as apples, bananas, and oranges. The agent's choice function formally represents which fruit(s) the agent would pick from any particular set of available ones, say from any particular fruit basket he or she is presented with. The function might look like this:

$$\begin{aligned} C(\{\text{apple,banana,orange}\}) &= \{\text{apple}\}; \\ C(\{\text{apple,banana}\}) &= \{\text{apple}\}; \\ C(\{\text{banana,orange}\}) &= \{\text{banana}\}; \\ C(\{\text{apple,orange}\}) &= \{\text{apple}\}; \\ C(\{\text{apple}\}) &= \{\text{apple}\}; \\ C(\{\text{banana}\}) &= \{\text{banana}\}; \\ C(\{\text{orange}\}) &= \{\text{orange}\}. \end{aligned}$$

That is, when all three fruits are available, the agent chooses an apple; when an apple and a banana are available, he or she also chooses an apple; when a banana and an orange are available, he or she chooses a banana; and so on. It is quite easy to explain the agent's choice function in this example: The agent simply prefers apples to bananas to oranges and always picks whichever fruit among the available ones is highest on this ranking.

Generally, any preference order $\succsim$ induces a corresponding choice function.[29] This is defined as follows:

$C(Y)=\{y \in Y : y \succsim x \text{ for all } x \in Y\}$ for any non-empty subset Y of X.[30]

[29] Conversely, any choice function that represents a sufficiently regular choice behaviour can be explained in terms of an underlying preference order or other binary relation. For a classic exposition of the relevant conditions, see Amartya Sen, 'Choice Functions and Revealed Preference', *Review of Economic Studies* 38 (3) (1971), pp. 307-317.

[30] To be well-defined, $C(Y)$ must always be non-empty. This requires that, for each subset Y of X, there is some alternative y in X that is maximal with respect to the preference order $\succsim$. This condition is trivially met if X is finite. It is also met if $\succsim=\succsim_M$, as defined in this paper, *and* M is a finite set of reasons, a plausible psychological hypothesis.

Thus we are immediately able to ascribe to any agent modelled by our theory not only a family of preference orders $\succsim_M$ across the different possible sets of motivating reasons M in $\mathcal{M}$, but also a family of corresponding choice functions C_M across all M in $\mathcal{M}$. Each such choice function C_M represents the agent's choice behaviour in the event that M is his or her set of motivating reasons.

The resulting picture of rational choice is fairly straightforward: At any given time, the agent is in a particular psychological state, represented by his or her set of motivating reasons, which, jointly with the agent's underlying generating relation, determines his or her preference order. This preference order now induces the agent's choice function, which tells us what the agent's choice behaviour would be when presented with any concrete set of alternatives. By implication, a change in the agent's set of motivating reasons can induce not only a change in his or her preference order, as we have seen, but also a change in the resulting choice function and thus in the agent's choice behaviour. Using the technical tools from the previous section, this picture can be further extended so as to incorporate choices under uncertainty as well, but for simplicity we set these technicalities aside here.

The picture just sketched, however, involves a simplifying assumption. By defining the agent's choice function C_M on the basis of the preference order $\succsim_M$, we have implicitly assumed that the agent's set of motivating reasons M is given independently of the particular choice situation to which that choice function is subsequently applied. Call this the case of an *exogenous* set of motivating reasons. We obtain a more sophisticated – and perhaps more realistic – picture of an agent's choice behaviour by allowing his or her set of motivating reasons to depend on the choice situation he or she finds him- or herself in. In other words, different choice situations may *endogenously* trigger different sets of motivating reasons.

Recall our example of an agent choosing between three different types of cars. If the agent's set of motivating reasons, which could be any subset of 'the car is fast' (F), 'it is big' (B), 'it is environmentally friendly' (E), were exogenously given, everything would be as in our example of choices

between fruits: The choice function over cars would look exactly like the one over fruits, just induced by the appropriate preference order over cars instead of the one over fruits. But if different choice situations somehow triggered different sets of motivating reasons, the agent's choice behaviour would be more complex. Suppose, for instance, that a choice between any two cars leads the agent to be motivated by all and only those reasons that distinguish those cars. The agent may then exhibit a 'cyclical' choice behaviour. In our example, he or she would choose the car that is fast and small ($f\neg be$) over the slow and big one ($\neg fbe$) when presented with two environmentally friendly cars, the slow and environmentally friendly car ($\neg fbe$) over the fast and environmentally unfriendly one ($fb\neg e$) when presented with two big cars, and the big and environmentally unfriendly car ($fb\neg e$) over the small and environmentally friendly one ($f\neg be$) when presented with two fast cars. The resulting choice function would not be explicable in terms of any single underlying preference order, since any such order would have to rank $f\neg be$ over $\neg fbe$, $\neg fbe$ over $fb\neg e$, and yet $fb\neg e$ over $f\neg be$, a violation of transitivity.

While standard rational choice theory lacks the ability to explain this kind of choice behaviour, our theory can account for it by pointing to the way in which the agent's set of motivating reasons is endogenously affected by the choice situation the agent is faced with. This kind of context dependence of preferences and choices is a widely observed phenomenon that is seen by many as a serious challenge to rational choice theory. By making rational choice theory reason-based as suggested in this paper, we have therefore introduced some new conceptual resources for analyzing this phenomenon.

13 Concluding remarks

We have proposed a reason-based theory of rational choice, which responds to the widely held concern that standard rational choice theory does not say anything about the reasons and motivations underlying preferences but holds preferences to be unchangeable and not subject to reason-based scrutiny. Our theory can be viewed from two angles. On the one hand, it generalizes

standard rational choice theory and thereby connects with the large body of work in decision theory and the social sciences on the formal modelling of human decision problems. On the other hand, it formalizes the role played by reasons in rational decision making, thereby capturing a core concern of the philosophical literature on the relationship between reasons and actions. Our theory should not be seen as a rival to either body of work. Instead, one of our aims is to initiate a dialogue between formal rational choice theory and philosophical work on reasons. Although we have only presented a first sketch of our theory and much further work is necessary, we hope that this paper will provide some of the concepts and tools needed to advance this enterprise.

A Appendix

We here prove theorems 1 and 2. The proofs of theorems 1* and 2* – the extensions to preferences over uncertain prospects – are significantly more technical and are available on request. Throughout our proofs, we write $M_x := \{R \in M : R \text{ is true of } x\}$ to denote the set of reasons in a given set $M \subseteq \mathcal{R}$ that are true of an alternative x in X, and we write $\mathcal{S} := \{S \subseteq \mathcal{R} : S \text{ is consistent}\}$ to denote the set of all consistent reason sets.

The proof of both theorems uses the following simple lemma, which holds independently of any assumptions on the set $\mathcal{M}$ of possible motivating sets.

Lemma 1. *Suppose axiom 1 holds. For all x, y, x', y' in X and all M in M, if $\{R \in M : R \text{ is true of } x\} = \{R \in M : R \text{ is true of } x'\}$ and $\{R \in M : R \text{ is true of } y\} = \{R \in M : R \text{ is true of } y'\}$ then $x \succsim_M y \Leftrightarrow x'_M \succsim y'$.*

Proof. Let $x, y, x', y' \in X$ and $M \in \mathcal{M}$ such that $M_x = M_{x'}$ and $M_y = M_{y'}$. Applying axiom 1 twice, we have $x \sim_M x'$ and $y \sim_M y'$. So, as $\succsim_M$ is transitive, $x \succsim_M y \Leftrightarrow x' \succsim_M y'$. ■

A.1 Proof of theorem 1

We first prove necessity and then sufficiency of our two axioms for the representation of preferences in terms of a generating relation.

Necessity of the axioms for the representation

First, suppose a binary relation $\geq$ on $\mathcal{S}$ generates all preference orders $\succsim_M$ across $M \in \mathcal{M}$. Axiom 2 is obviously satisfied. As for axiom 1, consider any $M \in \mathcal{M}$ and any $x, y \in X$ such that $M_x = M_y$. We have to show that $x \sim_M y$. As $\succsim_M$ is reflexive, we have $x \sim_M x$. So, since $\geq$ generates $\succsim_M$, we must have $M_x \equiv M_x$. But since $M_x = M_y$, this implies $M_x \equiv M_y$. From this – again using the fact that $\geq$ generates $\succsim_M$ – it follows that $x \sim_M y$, as required.

Sufficiency of the axioms for the representation

Now assume that axioms 1 and 2 are satisfied. Recall that $\mathcal{M}$ is closed under finite intersection. This is part (i) of our regularity assumption on $\mathcal{M}$. There is no need to assume part (ii) for Theorem 1.

Claim 1. *For all $x, y, x', y' \in X$ and all $M, M' \in \mathcal{M}$, if $M_x = M'_{x'}$ and $M_y = M'_{y'}$, then $x \succsim_M y \Leftrightarrow x' \succsim_{M'} y'$.*

To prove this claim, let $x, y, x', y' \in X$ and $M, M' \in \mathcal{M}$ with $M_x = M'_{x'}$ and $M_y = M'_{y'}$. As $\mathcal{M}$ is closed under finite intersection, we have $M \cap M' \in \mathcal{M}$. We first show that

$$(M \cap M')_x = (M \cap M')_{x'} = M_x = M'_{x'},$$
$$(M \cap M')_y = (M \cap M')_{y'} = M_y = M'_{y'}.$$

To see that the first set of identities holds, notice the following: firstly, $M_x = M'_{x'}$ by assumption; secondly, $(M \cap M')_x = M_x$, since $(M \cap M')_x = M_x \cap M'_x = M_x$ (the last identity holds because $M'_x \supseteq (M'_{x'})_x = (M_x)_x = M_x$); and, thirdly, $(M \cap M')_{x'} = M'_{x'}$, since $(M \cap M')_{x'} = M_{x'} \cap M'_{x'} = M'_{x'}$ (the last identity holds because $M_{x'} \supseteq (M_x)_{x'} = (M'_{x'})_{x'} = M'_{x'}$). The second set of identities holds by an analogous argument.

Now, since $(M \cap M')_x = M_x$ and $(M \cap M')_y = M_y$, axiom 2 implies

$$x \succsim_{M \cap M'} y \Leftrightarrow x \succsim_M y. \qquad (*)$$

Further, since $(M \cap M')_{x'} = M'_{x'}$ and $(M \cap M')_{y'} = M'_{y'}$, axiom 2 implies

$$x' \succsim_{M \cap M'} y' \Leftrightarrow x' \succsim_{M'} y'. \qquad (**)$$

Finally, since $(M \cap M')_x = (M \cap M')_{x'}$ and $(M \cap M')_y = (M \cap M')_{y'}$, lemma 1 implies

$$x \succsim_{M \cap M'} y \Leftrightarrow x' \succsim_{M \cap M'} y'. \qquad (***)$$

The equivalences (*) to (***) together imply that $x \succsim_M y \Leftrightarrow x' \succsim_{M'} y'$. ■

Claim 1 allows us to define a binary relation $\geq$ on $\mathcal{S}$ with the following properties: for all $S, S' \in \mathcal{S}$, $S \geq S'$ if and only if $x \succsim_M y$ for *some* (hence, by claim 1, *all*) $x, y \in X$ and $M \in \mathcal{M}$ such that $M_x = S$ and $M_y = S'$.

Claim 2. *For each $M \in \mathcal{M}$, $\geq$ generates $\succsim_M$, that is, $x \succsim_M y \Leftrightarrow M_x \geq M_y$ for all $x, y \in X$.*

To prove this claim, let $M \in \mathcal{M}$ and $x, y \in X$. First, assume $x \succsim_M y$. We show that $M_x \geq M_y$, that is, $x' \succsim_{M'} y'$ for some $x', y' \in X$ and $M' \in \mathcal{M}$ with $M'_{x'} = M_x$ and $M'_{y'} = M_y$. This obviously holds: simply take $x' = x$, $y' = y$, and $M' = M$. Conversely, assume that $M_x \geq M_y$. Then, by the definition of $\geq$ and claim 1, we have $x' \succsim_{M'} y'$ for *all* $x', y' \in X$ and $M' \in \mathcal{M}$ satisfying $M'_{x'} = M_x$ and $M'_{y'} = M_y$. In particular, $x \succsim_M y$. This completes the proof of theorem 1. ■

A.2 Proof of theorem 2

Assume a weakly independent set of reasons $\mathcal{R}$. The proof is written so as to maximize parallels with the proof of theorem 1.

Necessity of the axioms for the representation

By the argument in the earlier proof, axioms 1 and 2 hold if some order $\geq$ of consistent reason sets generates all preference orders $\succsim_M$, across $M \in \mathcal{M}$.

Sufficiency of the axioms for the representation

Now assume that axioms 1 and 2 are satisfied. Recall that, for any M, M' in $\mathcal{M}$, $\mathcal{M}$ contains some superset of $M \cup M'$. This is a weakened variant of part (ii) of our regularity assumption on $\mathcal{M}$. There is no need to assume part (i) for theorem 2.

Claim 1. *For all $x, y, x', y' \in X$ and all $M, M' \in \mathcal{M}$, if $M_x = M'_{x'}$ and $M_y = M'_{y'}$, then $x \succsim_M y \Leftrightarrow x' \succsim_{M'} y'$.*

This claim, although analogous to the first claim in the proof of theorem 1, requires a different proof. Let $x, y, x', y' \in X$ and $M, M' \in \mathcal{M}$ such that $M_x = M'_{x'}$ and $M_y = M'_{y'}$. As $\mathcal{M}$ contains M, M', it contains some $M'' \supseteq M \cup M'$, by assumption. By the weak independence of $\mathcal{R}$, there are

$a, b \in X$ such that $\mathcal{R}_a = M_x$ and $\mathcal{R}_b = M_y$. Hence $M''_a = M_x$ and $M''_b = M_y$, so that by axiom 2

$$a \succsim_{M''} b \Leftrightarrow a \succsim_M b. \qquad (*)$$

By an analogous argument (performed on x', y', M' instead of x, y, M), there are $a', b' \in X$ such that $M''_{a'} = M'_{x'}$ and $M''_{b'} = M'_{y'}$ and

$$a' \succsim_{M''} b' \Leftrightarrow a' \succsim_{M'} b'. \qquad (**)$$

Using lemma 1, the right-hand side of (*) is equivalent to $x \succsim_M y$ (because $M_a = M_x$ and $M_b = M_y$); the right-hand side of (**) is equivalent to $x' \succsim_{M'} y'$ (because $M'_{a'} = M'_{x'}$ and $M'_{b'} = M'_{y'}$); and the left-hand sides of (*) and (**) are equivalent to each other (because $M''_a = M''_{a'}$ and $M''_b = M''_{b'}$). These three equivalences together with the equivalences (*) and (**) imply that $x \succsim_M y \Leftrightarrow x' \succsim_{M'} y'$. ■

Claim 1 allows us to define a binary relation $\geq^*$ on $\mathcal{S}$, analogous to the one defined in our proof of theorem 1. But this time it is only a precursor to the relation we ultimately wish to define (it must subsequently be extended to an order). The relation $\geq^*$ has the following properties: for any $S, S' \in \mathcal{S}$, $S \geq^* S'$ if and only if $x \succsim_M y$ for *some* (hence, by claim 1, *all*) $x, y \in X$ and $M \in \mathcal{M}$ such that $M_x = S$ and $M_y = S'$.

Claim 2. *For each $M \in \mathcal{M}$, the binary relation $\geq^*$ generates $\succsim_M$, that is, $x \succsim_M y \Leftrightarrow M_x \geq M_y$ for all $x, y \in X$.*

The proof is analogous to that of the second claim in the proof of theorem 1. ■

Claim 3. *$\geq^*$ is transitive.*

Consider $S, S', S'' \in \mathcal{S}$ such that $S \geq^* S'$ and $S' \geq^* S''$. We have to show that $S \geq^* S''$. Since $S \geq^* S'$, there exist $x, y \in X$ and $M \in \mathcal{M}$ such that $M_x = S$, $M_y = S'$ and $x \succsim_M y$. Since $S' \geq^* S''$, there exist $y', z \in X$ and $M' \in \mathcal{M}$ such that $M'_{y'} = S'$, $M'_z = S''$ and $y' \succsim_{M'} z$. Since $M, M' \in \mathcal{M}$ and by our assumption on $\mathcal{M}$, $\mathcal{M}$ contains some $M'' \supseteq M \cup M'$. By the

weak independence of $\mathcal{R}$, there are $a, b, c \in X$ such that $\mathcal{R}_a = S$, $\mathcal{R}_b = S'$ and $\mathcal{R}_c = S''$, whence $M''_a = S$, $M''_b = S'$ and $M''_c = S''$. Since $x \succsim_M y$, $M_x = M''_a$ $(= S)$, and $M_y = M''_b$ $(= S')$, and by claim 1, we have $a \succsim_{M''} b$. Similarly, since $y' \succsim_{M'} z$, $M'_{y'} = M''_b$ $(= S')$, and $M'_z = M''_c$ $(= S'')$, and by claim 1, we have $b \succsim_{M''} c$. Since $a \succsim_{M''} b$ and $b \succsim_{M''} c$, and by the transitivity of $\succsim_{M''}$, we have $a \succsim_{M''} c$. So, by the definition of $\geq^*$ (and using the fact that $M''_a = S$ and $M''_c = S''$), we have $S \geq^* S''$. ■

Claim 4. *There exists an order $\geq$ on $\mathcal{S}$ that extends $\geq^*$, in the usual sense that $S >^* S' \Rightarrow S > S'$ and $S \equiv^* S' \Rightarrow S \equiv S'$ for all $S, S' \in \mathcal{S}$; equivalently, $S \geq S' \Leftrightarrow S \geq^* S'$ for all $S, S' \in \mathcal{S}$ that are ranked relative to each other by $\geq^*$.*

This follows from claim 3 via a classic extension theorem for binary relations.[31] ■

Claim 5. *For each $M \in \mathcal{M}$, the order $\geq$ defined in claim 4 generates $\succsim_M$.*

To prove this claim, let $M \in \mathcal{M}$ and $x, y \in X$. First, if $x \succsim_M y$, then $M_x \succsim^* M_y$, as $\geq^*$ generates $\succsim_M$ (by claim 2), whence $M_x \geq M_y$, as $\geq$ extends $\geq^*$. Conversely, if $x \not\succsim_M y$, then $y \succ_M x$ (as $\succsim_M$ is complete), so that $M_y >^* M_x$, as $\geq^*$ generates $\succsim_M$ (by claim 2). This implies that $M_y > M_x$, since $\geq$ extends $\geq^*$, hence that $M_x \not\geq M_y$. This completes the proof of theorem 2. ■

[31] As proven in its most general form by K. Suzumura, 'Remarks on the theory of collective choice', *Economica* 43 (1976), pp. 381–90.